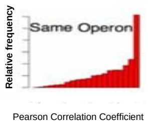
Combining Comparative Genomics and Expression Data to Predict Regulatory Networks

Richard Koche[1], Morgan Price[1], Eric Alm[1], Adam Arkin[1,2]

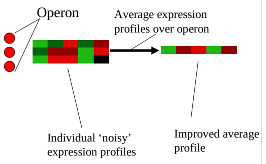
[1]Division of Physical Sciences, Lawrence Berkeley National Lab; [2]Department of Bioengineering, University of California, Berkeley

Abstract. Genome organization in prokaryotes has been well-studied, and gene location alone can often be used to infer functional relationships. We combine high-quality operon predictions with comparative genomics to identify likely sets of co-regulated genes. We use this *a priori* biological knowledge to improve traditional gene expression profile clustering techniques.

Background. The most direct way to control the expression of a gene is to regulate its rate of transcription. For functionally-coupled genes whose expression needs to be coordinated, this transcriptional control is often accomplished by placing them on the same polycistronic mRNA, or operon.



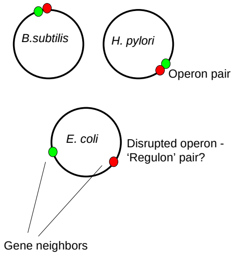
Clustering gene expression profiles. Clustering gene expression data is one of the most common approaches to identifying functionally related genes. However, this method can be improved upon if *a priori* relationships among a group of genes are known. For genes co-transcribed in operons, high-quality operon predictions can be used to elucidate co-expression relationships. Then, these often noisy expression profiles can be averaged over all the genes in each operon, maximizing the signal-to-noise ratio. Such is the case with operons. Knowledge can be further extended by examining pairs of genes across all genomes to include operons which may have been broken up during evolution.

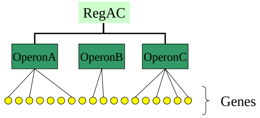

Operon
Average expression profiles over operon
Individual 'noisy' expression profiles
Improved average profile

Method. 1) Positional Clustering

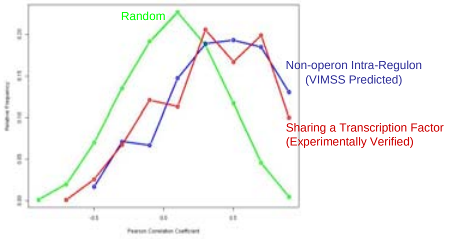
Gene Neighbor Method—assume gene clusters from phylogenetically distant organisms imply co-regulation (Overbeek *et. al.* 1999)

- Conserved gene clusters in prokaryotes are often composed of functionally related genes
- More genomes = higher confidence (currently 165 genomes in VIMSS database)
- The further the phylogenetic distance between 2 organisms, the less likely a cluster is due to chance





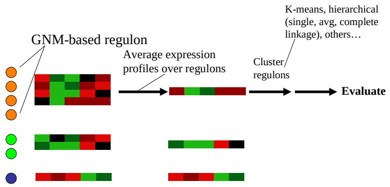
Building regulons from operons. Operons that appear near each other frequently enough in disparate genomes get combined into regulons: all pairwise GNM scores are computed, then averaged over operons to give operon-operon distances. These distances are then used cluster operons into regulons by using complete linkage hierarchical clustering. Thus, an operon is merged into a regulon cluster if it has at least one gene that is a "gene neighbor" with each of the other operons in the cluster .



Disrupted Operons Maintain Similar Promoters. Genes found within position-based regulons show a high expression correlation, even with the operon relationships removed. Strikingly, the distribution is statistically indistinguishable from genes known to share a transcription factor*. (Kolgomorov-Smirnov Test, p-value = 0.34, D = 0.0641, where D=0 for identical distributions and D=1 is for non-overlapping distributions).

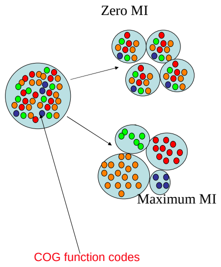
*DPInteract: http://arep.med.harvard.edu/dpinteract/

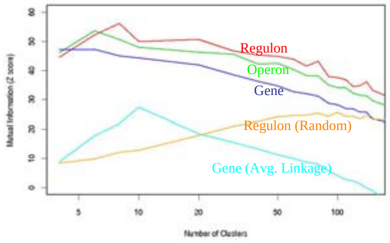
2) Regulon-regulon Coexpression: Clustering Algorithm


K-means, hierarchical (single, avg, complete linkage), others…
GNM-based regulon
Average expression profiles over regulons
Cluster regulons
Evaluate

Evaluating Clustering Results. We use mutual information of gene functions and clusters to measure effectiveness. Higher mutual information (MI) indicates a more homogeneous set of functions within the cluster. The COG functional categories are used to assess gene function. Because the scale of MI can be somewhat arbitrary, we compare observed MI to that for randomly generated clusters (of uniform size) with the same gene set, and compute a Z-score indicating functional enrichment above random.

Z-score = $(Mi_{obs} - \langle Mi_{random} \rangle)/\sigma(Mi_{random})$



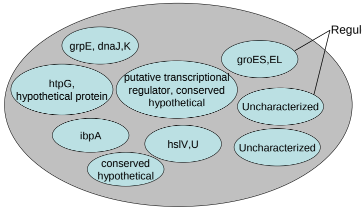


Results. Clustering of both operons and regulons consistently outperformed traditional gene-based methods as measured by mutual information of gene function and cluster. Further, regulons were able to improve performance even at low cluster numbers, and outperformed clustering of regulons with randomized expression profiles. Taken together, these results suggest that pre-grouping genes into regulons improved the ability of the clustering algorithms to utilize the microarray data.

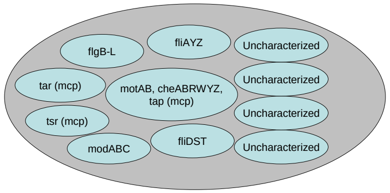
All clustering was done using the K-means algorithm, with the exception of average linkage hierarchical clustering, as indicated above. While in common use, this type of hierarchical clustering often gives worse-than-random results (Gibbons *et. al.* 2002). Curves above are the result of averaging 10 independent runs of the K-means clustering algorithm with different random seeds.

"Zooming in" on regulon clusters illuminates biologically relevant interactions, and often lends insight into potential relationships with as yet uncharacterized genes as shown in the examples below.

S. oneidensis Heat Shock Regulon
(Expression Data Courtesy of the Zhou lab)



E. coli Chemotaxis Regulon
(Expression Data Courtesy of ASAP database Blattner *et. al.*)


fliAYZ
flgB-L
Uncharacterized
tar (mcp)
motAB, cheABRWYZ, tap (mcp)
Uncharacterized
tsr (mcp)
Uncharacterized
modABC
fliDST
Uncharacterized